Precise Measurement of Position and Attitude Based on Convolutional Neural Network and Visual Correspondence Relationship

Jiachen Yang , Member, IEEE,JiabaoMan , Meng Xi, Xinbo Gao , Senior Member, IEEE, Wen Lu, Member, IEEE, and Qinggang Meng , Senior Member, IEEE

***Abstract*— Accurate measurement of position and attitude information is particularly important. Traditional measurement methods generally require high-precision measurement equipment for analysis, leading to high costs and limited applicability. Vision-based measurement schemes need to solve complex visual relationships. With the extensive development of neural networks in related fields, it has become possible to apply them to the object position and attitude. In this paper, we propose an object pose measurement scheme based on convolutional neural network and we have successfully implemented end-to-end position and attitude detection. Furthermore, to effectively expand the measurement range and reduce the number of training samples, we demonstrated the independence of objects in each dimension and proposed subadded training programs. At the same time, we generated generating image encoder to guarantee the detection performance of the training model in practical applications.**

***Index Terms*— Convolutional neural network, generating image encoder, position and attitude measurement, subadded picture.**

I. Introduction

THE accurate acquisition of real-time object and attitude information is critical for a wide range of applications. Based on radar, laser, and other pose measurement sensors, high requirements are placed on such measurement equipment. Moreover, most of these solutions require strict measurement conditions and cumbersome measurement procedures. These problems pose can present limitations to acquiring position and posture information.

With the development of computer technology, monocular vision measurement is widely used in different fields such as aviation, navigation, aerospace, industry, and military [1]. Compared to traditional inertial device-based measurement methods, monocular vision pose measurement has the advantage of not requiring physical contact [2], [3]. In addition, this measurement method is not subject to the measurement of the internal structure of the measured object, reducing the complexity of measurement equipment. However, such vision-based measurements need to determine the correspondence between specific feature points on the measured object and their image pixel locations.

Several studies have investigated pose measurement algorithms, features for pose measurement, 3-D reconstruction, and the test environment for pose measurement [4]–[9]. For instance, [10] and [11] described the composition of the pose measuring system in video guidance sensor (VGS). A physical map of multiple path beyond line of sight communication was introduced in [12] and [13]. The feature points in the measurement were extracted using the foreground image minus the background image. There was a filter film on the target to be measured, and the camera was equipped with a laser lighting device. However, the use of laser measuring equipment and filter films has increased the difficulty of applying such methods in a general-purpose environment. Because of this, it is particularly necessary to find a smart and efficient measurement scheme with high measurement accuracy, low time consumption, and simple equipment requirements.

We propose a real-time position-and-attitude information solution for objects, which is based on a convolutional neural network [14]. This scheme uses the monocular vision system to obtain a specific marker image on the target object.

In end-to-end validation experiments, by preprocessing the acquired images and putting them into the neural network, the well-trained pose measurement model is obtained to achieve an accurate measurement.

Furthermore, to reduce the number of neural network training sets, a wide range of angle and attitude information measurements is simultaneously achieved. We demonstrate the independence of pose movement in each dimension of the object in the pixel coordinate system, so that the excellent model training can be achieved without large traversal of all measurement angles. Using visual independence, we created a new training image—subadded picture. This kind of picture is used for network training. However, there is an absolute

J. Yang, J. Man, and M. Xi are with the School of Electronic Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: yangjiachen@tju.edu.cn; manjiabao@tju.edu.cn; ximeng@tju.edu.cn).

X. Gao is with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi'an 710126, China (e-mail: xbgao@mail.xidian.edu.cn).

W. Lu is with the School of Electronic Engineering, Xidian University, Xi'an 710126, China (e-mail: luwen@mail.xidian.edu.cn).

Q. Meng is with the Department of Computer Science, Loughborough University, Loughborough LE11 3TU, U.K. (e-mail: q.meng@lboro.ac.uk).

difference between subadded picture and the actual image we get. With the idea of generating network [15], we use the encoder network [16] to achieve the conversion of captured images and training images.

The proposed theory is tested on its own data set. The experimental results show that our algorithm has outperformed the current state-of-the-art results, in which the existing vision-based position and attitude measurement solutions.

The main contributions of this paper can be summarized in the following three points.

1) To the best of authors' knowledge, this paper is the first to apply convolutional neural networks to the continuous accurate measurement of position and attitude. Our main contribution is a rigorous evaluation of multiple CNN models and factors. By comparing the results with other traditional methods, it illustrates how the CNN-based methods represent the state of the art of position and pose estimation.
2) We theoretically analyze the visual correspondences between dimensions in the process of position and attitude movement, which groundbreakingly demonstrate the visual independence of each dimensional change. This theory lays a solid foundation for effectively reducing the number of neural network training sets. Through the effective preprocessing of pictures, a small picture training set is made to obtain an excellent test model. At the same time, the theory has also expanded the range of angle measurements and enhanced the effectiveness of analysis.
3) Based on the above-mentioned approach, we introduce the theory of generative networks and present it as an image encoder. Realizing the effective transformation between the actual project picture and the training picture, it enhances the universal applicability of the program in more complex situations.

II. Related Work

Precise measurement of object position and attitude has been investigated from different points of view and with different techniques. Among them, the monocular vision pose measurement system is widely used in various fields. It has the advantages of simple structure, low cost, and strong real-time performance. However, existing monocular vision measurement systems all require artificial calibration of image models, camera internal and external parameter models, and lens distortion calibration models. These tasks can be automated through machine learning related theories. We first effectively preprocess the feature marker image, which is based on the visual correspondence of the object pose movement. By using convolutional neural networks, the high-level abstraction of the processed image is performed. At the same time, we achieved the efficient conversion of training images and actual images through a generative encoder. Our work is to automate the calibration of complex models in visual correspondence. The measurement range of the sample is enlarged, the measurement accuracy is improved, and the real-time measurement performance is enhanced [17].

A. Visual Correspondence Analysis

The main modeling objects in machine vision are measurement cameras and targets. The focus of the current pose measurement includes the transformation between coordinate systems and camera internal and external parameter calibrations. At the same time, it also involves the extraction of target features, feature matching, pose calculation algorithm, and other aspects in the image. In the analysis of the image coordinate system, Gonzalez *et al.* [18] proposed the relationship between the image pixel system and the image physical coordinate system and the conversion rule. Szeliski [19] proposed an aperture imaging model that best fits the actual imaging situation for the camera imaging model. The weak perspective imaging model proposed by Gold *et al.* [20] effectively simplifies the small hole imaging model. At the same time, in the actual measurement, it is often necessary to convert the camera's measurement results to the world coordinate system. For example, the measurement results are converted to the robot coordinate system in the robot system [21]. The aforementioned helps the system make decisions and contributes to analyzing the accuracy of the measurement system's measurement results. In addition, the camera parameters need to be calibrated before taking measurements [22]. At the same time, it also needs to consider the problem of lens distortion. Otherwise, it will affect the measurement performance [23].

In the feature matching process, the existing method usually uses the difference method to perform feature extraction of the cooperation target. This method is implemented using laser illumination. With a specific wavelength laser, two types of images are taken: foreground and background images. Subtracting the two images yields a simple image containing the target.

Human-made calibration and extraction will inevitably cause a lot of loss of precision, and the time cost is high. Because of this, we rely entirely on machine learning for the above process, which can effectively avoid the above problems. By constructing an efficient neural network, useful training samples are made to implement the iteration and update of internal parameters. Furthermore, the correspondence between the camera and the specific marker can be automatically determined.

B. Position and Attitude Measurement

Measurement schemes based on cooperation goals or noncooperative goals are the two main approaches to the current position-and-attitude information solving. Heaton and Howard [24] conducted a detailed analysis of the measurement of cooperation objectives. Such a scheme can accurately extract the centroid of the feature point on the image and reduce the extraction error. Noncooperative target systems do not require specially designed cooperation goals. Therefore, feature extraction is a crucial step in the noncooperative target pose measurement, and it has the characteristics of high difficulty in the extraction and low extraction accuracy [25], [26].

A measurement scheme with a defined cooperation goal can significantly enhance measurement universality [27]. For this reason, we use the form of cooperation objectives in the program for adequate training and testing.

C. Conversion and Detection Based on Convolutional Neural Networks

In recent years, convolutional neural networks have played a key role in all areas of pattern recognition. Such networks have achieved significant results in the detection, classification, and regression based on image analysis [28]–[32]. Krizhevsky *et al.* [14] proposed AlexNet. The data expansion strategy proposed in this paper has solved the problems in data training. Afterward, Zeiger and Fergus [33] proposed ZF Net, Simonyan and Zisserman [34] proposed VGG Net, and Szegedy *et al.* [35] proposed GoogLeNet further improved and optimized it.

In [36], ResNet, proposed by Microsoft Research Asia in 2015, effectively solves the convergence problem of deep networks. Based on this, the network is widely used in various fields such as detection, segmentation, and identification. In this paper, we try to use this type of system for active position and attitude detection. At the same time, related work is also dedicated to exploring the impact of different depth network layers on measurement results. Experiments have achieved excellent results. To better solve the problem of gradient disappearance, DenseNet, proposed by Huang *et al.* [37] directly connects all layers of the network. As already stated, this effectively reduces the network parameters and width and makes full use of the feature map. We use this method to make full use of the information in the feature map. So that when the physical information of the object is minimal, the network can still better distinguish the tiny difference between the picture's pixels.

Goodfellow *et al.* [15] proposed a generative confrontation network, which effectively improved the efficiency of unsupervised learning. This network is widely used for image generation, such as superresolution tasks, semantic segmentation, etc., in a complex distributed environment. We expertly combine this network with a picture encoder. This approach aims to establish a picture conversion and generation system, and the useful link between the training picture and the actual picture is achieved finally. The specific application of this theory in this article will be discussed in detail in the next chapter.

III. Position and Attitude Measurement Based on Convolutional Neural Network and Generating Image Encoder

This new position and attitude measurement method will be discussed in the following two aspects. First, we obtained images of specific markers with a different position and posture information, and they were effectively preprocessed. We use labeled convolutional neural networks for labeling training. Furthermore, to improve measurement accuracy and reduce the amount of training required, we analyzed the visual correspondence between six dimensions of information change [38]. Based on this, a sufficient and minimal number of training sets is realized, and the conversion between test images and training sets is completed.

A. End-to-End Information Detection

The image is a quantitative representation of the real scene by the camera, transformed from the real scene to the image through the following processes.

1) The conversion of the object from the world coordinate system to the camera coordinate system.
2) Perspective imaging of the camera.
3) After imaging, the transformation from the picture physical's coordinate system to the picture pixel coordinate system.

Therefore, the image can be used to calculate the position and posture relationship of the object in the real scene.

Spatial position transformation is the basis of visual theory. Many experts and scholars have made research on coordinate system transformation between the two rigid body coordinate systems [39].

In general, the distance between the camera and the target is much larger than the target size when acquiring and photographing the feature marker. Based on this feature, it can be assumed that the distance from all points to the camera is a constant. At the same time, the imaging model of the camera is modeled using a small hole model.

Finally, the image physical coordinate system uses the center of the image as the origin, and the image pixel coordinate system uses the upper left corner of the image as the origin. Since the image pixel coordinate system is used in machine vision processing, the point on the image needs to be converted.

We define the world coordinate of the point as (X, Y, Z), the picture pixel coordinate as (u, v), and the camera focal length as f. The translation and rotation between the two rigid bodies are represented by vectors T and R. We assume that the dimensions of the charge-coupled device (CCD) pixels are d_x and d_y

$$Z_C \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{1}{dx} & 0 & 0 \\ 0 & \frac{1}{dy} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \quad (1)$$

The matrix contains a total of 11 parameters that can be calibrated by the camera. Define a matrix M that represents the camera's internal and external parameters

$$M = \begin{bmatrix} \frac{1}{dx} & 0 & 0 \\ 0 & \frac{1}{dy} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \\ m_{41} & m_{42} & m_{43} & m_{44} \end{bmatrix}. \quad (2)$$

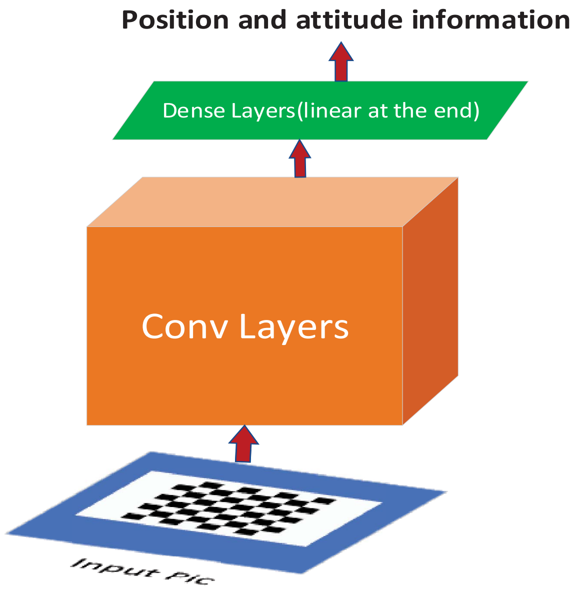

Position and attitude information
Dense Layers(linear at the end)
Conv Layers
Input Pic

Fig. 1. End-to-end position and attitude detection system architecture.

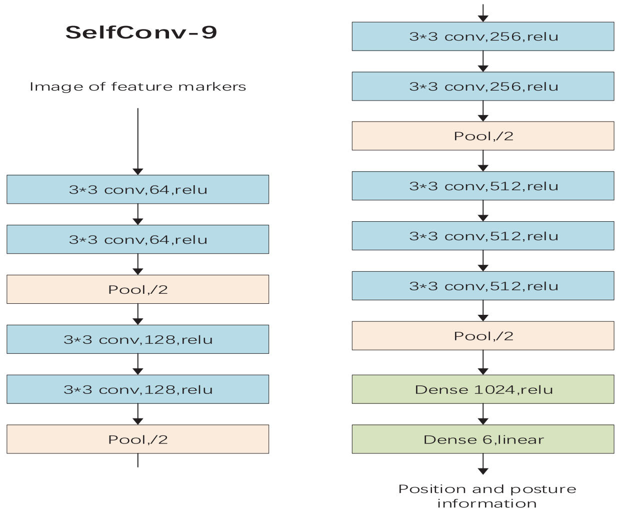


Fig. 2. CNN structure for object position and attitude information detection.

In view of the above-mentioned visual correspondence analysis, there is a specific relationship between position and attitude movement and picture pixel distribution. We used convolutional neural networks to create feature extractors consisting of convolutional layers and subsampling layers, and use it to extract pixel distribution features. The information regression analysis is performed by changing the final layer activation function of the network to be linear.

According to the previously described visual correspondence theory, convolutional networks can be used to perform the regression analysis of position and attitude information. In the end-to-end detection demonstration experiment, we built a shallow convolutional network for testing, with nine convolution layers.

Experiments used black-and-white checkerboard patterns as feature markers. We have obtained a checkerboard picture in different poses and positions and assigned it to a set of six dimensions to make it a training set. Based on this, we complete the neural network training.

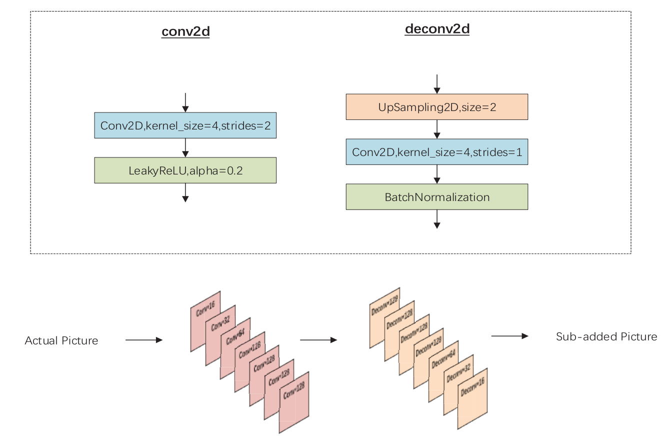


Fig. 3. Generating encoder network structure.

To make the training effect better, we have reduced the sample acquisition interval in each dimension. The sample minimum acquisition interval is controlled to around 0.01 (i.e., position information is not less than 0.01 m, attitude information is not less than 0.01°).

Fine sample acquisition is helpful to enrich the feature representation in different information changing situations. The aforementioned can enhance the feature extraction accuracy of convolutional neural networks. At the same time, we are also aware that with a wide range of attitude and distance information acquisition, this fine sample acquisition will result in an exponential increase in the number of necessary training sets. For both, we have conducted a deeper exploration of visual correspondence. The already stated can not only significantly reduce the number of samples but can also guarantee the training accuracy of neural networks. Fig. 1 shows the detection architecture of the system. Fig. 2 provides a detailed description of the CNN network structure we have defined with good detection accuracy.

B. Subadded Picture

Analyzing the single feature point on the cooperative target, we can see that in the process of continuously shooting two images by the camera, the target only translates Δx in the X-axis direction, and then the feature point also translates Δx in the X-axis direction. Let the feature point on the cooperation target be P, then the space coordinate of point P will change from (x_0, y_0, z_0) to P' $(x_0 + \Delta x, y_0, z_0)$. We perform image processing on the two consecutive images taken by the camera and feature point matching. The pixel coordinate of the feature point P on the first image is (u_0, v_0), and the pixel coordinate of P' on the image is $({u_0}', {v_0}')$. Subtracting pixel coordinates, we can get the following formula:

$$\begin{aligned} Z_c' \begin{bmatrix} {u_0}' \\ {v_0}' \\ 1 \end{bmatrix} - Z_c \begin{bmatrix} u_0 \\ v_0 \\ 1 \end{bmatrix} &= M \begin{bmatrix} x_0 + \Delta x_0 \\ y_0 \\ z_0 \\ 1 \end{bmatrix} - M \begin{bmatrix} x_0 \\ y \\ z \\ 1 \end{bmatrix} \\ &= \Delta x \begin{bmatrix} m_{11} \\ m_{21} \\ m_{31} \end{bmatrix}. \end{aligned} \tag{3}$$

In the same way, if the target is only along the Y-axis and the Z-axis is shifted by Δy and Δz, we can get the following relationship:

$$Z_c' \begin{bmatrix} u_0{}' \\ v_0{}' \\ 1 \end{bmatrix} - Z_c \begin{bmatrix} u_0 \\ v_0 \\ 1 \end{bmatrix} = M \begin{bmatrix} x_0 \\ y_0 + \Delta y_0 \\ z_0 \\ 1 \end{bmatrix} - M \begin{bmatrix} x \\ y_0 \\ z \\ 1 \end{bmatrix} = \Delta x \begin{bmatrix} m_{12} \\ m_{22} \\ m_{32} \end{bmatrix} \tag{4}$$

$$Z_c' \begin{bmatrix} u_0{}' \\ v_0{}' \\ 1 \end{bmatrix} - Z_c \begin{bmatrix} u_0 \\ v_0 \\ 1 \end{bmatrix} = M \begin{bmatrix} x_0 \\ y_0 \\ z_0 + \Delta z_0 \\ 1 \end{bmatrix} - M \begin{bmatrix} x \\ y \\ z_0 \\ 1 \end{bmatrix} = \Delta x \begin{bmatrix} m_{13} \\ m_{23} \\ m_{33} \end{bmatrix}. \tag{5}$$

It can be inferred that the spatial translation of the target can be reflected in the picture pixel coordinate system. Also, there is a linear relationship between the translational motion and the camera's internal and external parameters. When the camera continuously shoots two images, if the target rotates θ_z only around the Z-axis, the feature point also rotates only θ_z about the Z-axis. As above, subtracting the pixel coordinates of the feature points in the two images, we can get

$$\begin{aligned} Z_c' = \begin{bmatrix} u_0{}' \\ v_0{}' \\ 1 \end{bmatrix} - Z_c \begin{bmatrix} u_0 \\ v_0 \\ 1 \end{bmatrix} = M \begin{bmatrix} \cos\theta_z x_0 + \sin\theta_z y_0 \\ -\sin\theta_z x_0 + \cos\theta_z y_0 \\ z_0 \\ 1 \end{bmatrix} \\ - M \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = [(\cos\theta_z - 1)x_0 + \sin\theta_z y_0] \begin{bmatrix} m_{11} \\ m_{21} \\ m_{31} \end{bmatrix} \\ + [-\sin\theta_z x_0 + (\cos\theta_z - 1)y_0] \begin{bmatrix} m_{12} \\ m_{22} \\ m_{32} \end{bmatrix}. \end{aligned} \tag{6}$$

Similarly, if the target is only around the Y-axis and the X-axis is rotated by θ_y and θ_x, then

$$\begin{aligned} Z_c' = \begin{bmatrix} u_0{}' \\ v_0{}' \\ 1 \end{bmatrix} - Z_c \begin{bmatrix} u_0 v_0 \\ 1 \end{bmatrix} = M \begin{bmatrix} \cos\theta_y x_0 + \sin\theta_y z_0 \\ y_0 \\ -\sin\theta_y x_0 + \cos\theta_y z_0 \\ 1 \end{bmatrix} \\ - M \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = [(\cos\theta_y - 1)x_0 + \sin\theta_y z_0] \begin{bmatrix} m_{11} \\ m_{21} \\ m_{31} \end{bmatrix} \\ + [-\sin\theta_y x_0 + (\cos\theta_y - 1)z_0] \begin{bmatrix} m_{13} \\ m_{23} \\ m_{33} \end{bmatrix} \end{aligned} \tag{7}$$

$$\begin{aligned} Z_c' = \begin{bmatrix} u_0{}' \\ v_0{}' \\ 1 \end{bmatrix} - Z_c \begin{bmatrix} u_0 \\ v_0 \\ 1 \end{bmatrix} = M \begin{bmatrix} x_0 \\ \cos\theta_x y_0 + \sin\theta_x z_0 \\ -\sin\theta_x y_0 + \cos\theta_x z_0 \\ 1 \end{bmatrix} \\ - M \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = [(\cos\theta_x - 1)y_0 + \sin\theta_x z_0] \begin{bmatrix} m_{12} \\ m_{22} \\ m_{32} \end{bmatrix} \\ + [-\sin\theta_x y_0 + (\cos\theta_x - 1)z_0] \begin{bmatrix} m_{13} \\ m_{23} \\ m_{33} \end{bmatrix}. \end{aligned} \tag{8}$$

Therefore, it can be inferred that the spatial rotation of the target can be reflected in the picture pixel coordinate system. Also, there is a linear relationship between the rotational motion and the camera's internal and external parameters.

Because of this, we propose to use the sum of independent mobile information in each dimension of the object instead of the six dimensions to jointly move information. We use this as a training sample to replace the actual required joint movement information.[1] To ensure adequate information acquisition, we perform binarization processing on all images obtained. The aforementioned can mostly avoid the influence of light, rain, fog, and other weather on the visual information of the signature. In addition, we also perform other effective preprocessing on the actual samples obtained. Before the sample is put into training, the binarized information-containing picture is subtracted from the reference picture to form a difference map. The already stated allows neural networks to extract information changes more efficiently. This change implies the vectorial nature of the position and attitude movements. In general, the training samples that have been effectively preprocessed should be compared with the baseline image first, and then the individual moving images in each dimension should be added. We term it as the picture based on the above method subadded picture.

Based on this, we have achieved the goal of accomplishing large-scale training missions using very few training sets. In the case of determining the angular measurement range, the measured pace is adjusted in each dimension, and sample acquisition of independent information is completed, and through the combination of ways to eventually achieve the construction of a large-scale training set.

C. Generating Image Encoder

Subadded picture has an accurate of 6-D position and attitude information, but it is different from the visual image obtained in the actual project. However, both types of images contain the same vector change relationship so that they can be transformed into each other. Given this, we combine the generative network architecture with the image encoder with

[1] For example, we need to obtain a sample with six dimensions of $(m_1, m_2, m_3, m_4, m_5, m_6)$. At this point, we can create $(m_1, m_0, m_0, m_0, m_0, m_0)$, $(m_0, m_2, m_0, m_0, m_0, m_0)$, $(m_0, m_0, m_3, m_0, m_0, m_0)$, $(m_0, m_0, m_0, m_4, m_0, m_0)$, $(m_0, m_0, m_0, m_0, m_5, m_0)$, $(m_0, m_0, m_0, m_0, m_5, m_6)$, a total of six types of independent mobile information pictures, and add and process the six types of picture information at the pixel level. m_0 represents the reference value in each information dimension.

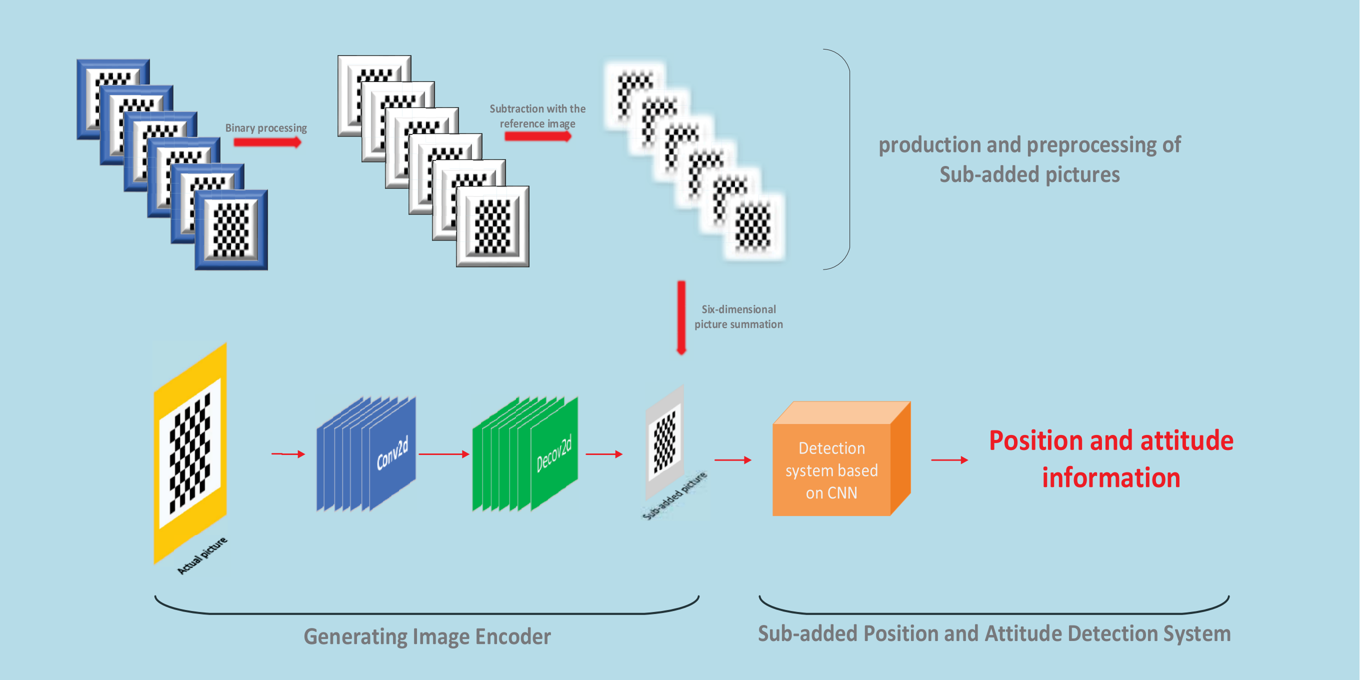

Binary processing
Subtraction with the reference image
production and preprocessing of Sub-added pictures
Six-dimensional picture summation
Actual picture
Conv2d
Decov2d
Sub-added picture
Detection system based on CNN
Position and attitude information
Generating Image Encoder
Sub-added Position and Attitude Detection System

Fig. 4. Position and attitude detection framework based on subadded pictures and generating image encoder.

TABLE I
POSITION AND ATTITUDE TEST ERROR IN END-TO-END EXPERIMENTS

Method	Positive Distance (m)	Lateral Offset (m)	Vertical Offset (m)	Roll Angle (”)	Pitch Angle (”)	Rotation Angle (”)
NASA-AVGS	0.12	0.12	0.12	720	720	468
NASA-NGAVGS	0.013	0.013	0.076	540	540	1224
DARPA-OE	0.12	0.12	0.12	720	720	468
ETS-VII	0.025	0.025	0.025	468	468	468
OAC	0.0432	0.0158	0.0012	428	23.4	361
Our Method	**0.00606**	**0.00186**	**0.00149**	**25.154**	**19.874**	**9.758**

We conducted performance tests in six dimensions including position and attitude. Each parameter in the table is the absolute value obtained by subtracting the actual measured value from the theoretical value. The first four rows of the table compare the relative measurement accuracy of the existing position and attitude measurement solutions. It can be seen that the end-to-end measurement scheme based on the convolutional neural network has obtained good measurement results. In six dimensions, its performance is significantly superior to other traditional types of measurement programs.

–**ETS-VII** [40] is a two-way remote operation test platform built by the Japan Space Agency, and it has been subjected to engineering rendezvous and docking test.

–The **Orbital Express (OE)** [41] undertakes the engineering tasks of automatic meeting and docking of space. Test ASTRO and NEXTSat docking on track and a variety of service operations.

–The **advanced VGS (AVGS)** [42] is a sensor mounted on the OE to provide the relative position and attitude between ASTRO and NEXTSat.

–The **next generation AVGS** [43] further improved the technical indicators based on AVGS.

–The **OAC** [44] is a six-dimensional information measurement algorithm based on laser and monocular vision.

the help of the theory of generative adversarial networks. Generating image encoder comes from this.

We adjusted the original image encoder architecture. The visual image acquired by the actual project is used as an encoder input, and the subadded picture used for network training is output as an encoder. In this way, the training model is applied in practical engineering. Fig. 3 provides a detailed description of the internal architecture of Generating Image Encoder. Fig. 4 shows the global architecture of the detection network based on Subadded Pictures and Generating Image Encoder.

IV. EXPERIMENTS

Based on the 3Dmax model making software, we acquired 117 649 actual marker images with different locations and information and made them into a training set. The experiment was performed on the data set we made. We chose black-and-white checkerboards as feature markers. Within the measurement interval, several such images were taken for neural network training. The 6-D information corresponding to these images is accurate to two decimal places.

A. Experiments of End-to-End Inspection

We developed our model by using the python deep learning library Keras [45] on a PC with a single 3.2-GHz CPU and a single GTX1080 GPU.

In the end-to-end verification stage of verification, we chose the ResNet–50 as the detection network architecture. In the measurement interval, we photographed several of these

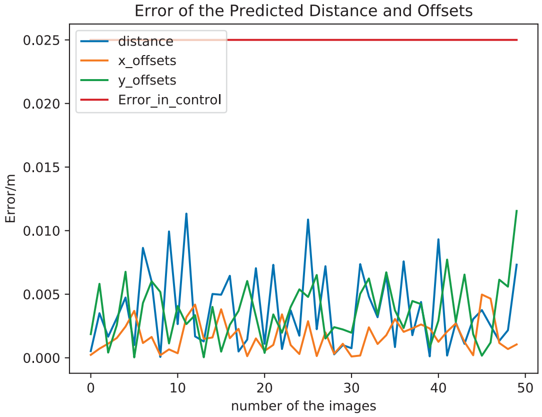


Fig. 5. Error of the predicted distance and offsets.

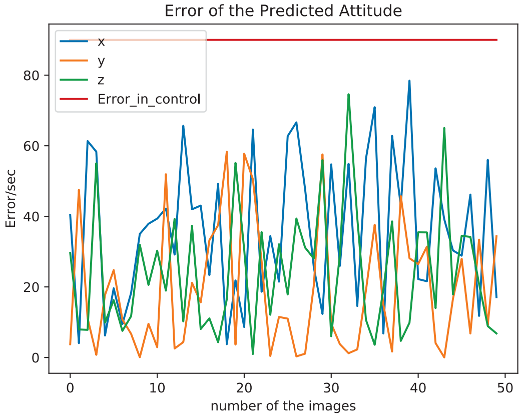


Fig. 6. Error of the predicted attitude.

images for training on the Internet. The number of full connections in these network layers is gradually attenuated. Based on the above visual correspondence, the attitude and position information has a specific relationship with the pixel point distribution of the image. Therefore, we adjust the final layer activation function to *linear*, the number of full-connection final layers is set as 6 [46]. With this setting, all six dimensions of information could be accurately output.

Based on the Kreas neural network framework, combined with the storage performance of existing equipment, we performed model iterative training. During the training of each model, the network randomly captures 2601 pictures in the training set to form a group of images. The epoch control for each model is 2, the batch size is set to 2, and the learning rate is set to 0.03. After the current model is trained and saved, the network will first read the relevant weight information of the previous model, and then randomly grab the training set image for the next round of training. We conducted a total of 50 such model iteration training. The network training effect was positively related to the scale and accuracy of the training set. We first performed an ergodic sample acquisition within the measured interval. The sample acquisition step was 0.01. We believe that such effective training set coverage contributes to the generalization of the network and enhance its robustness. Table I shows the accuracy of our model in the detection of position and attitude information.

Analyzing the relevant information in the table, we can see that ergodic sample acquisition helps to enhance the generalization of the network. The aforementioned can significantly improve the test accuracy of the model. However, the exponential growth of the number of samples caused by this has also challenged the acquisition and training of samples.

Figs. 5 and 6 show a visual analysis of the position and attitude errors, respectively. We set the error control threshold line separately on the two types of information dimensions involving position and attitude. Among them, the position-related threshold line was set to 0.025 m, and the gesture line was set to 90 s. It can be seen that 100% of the test samples are controlled within the threshold line.

The above-mentioned experimental results demonstrate the end-to-end detection performance based on convolutional neural networks. The implementation of high-precision detection systems also depends on the acquisition of traversal samples. However, the difficulty of traversing a sample is directly proportional to the detection range. Once the scope of information detection is expanded, the acquisition of such samples will certainly become an engineering difficulty. Table II shows the effect of sample coverage on position and attitude measurements in end-to-end experiments. Given this, we have reduced the size of the training sets and conducted experimental tests.

TABLE II
EFFECT OF SAMPLE COVERAGE ON POSITION AND ATTITUDE MEASUREMENTS IN END-TO-END EXPERIMENTS

Sample Coverage Ratio	Global Mean Square Error	Global Average Error
1	0.000026851	0.00251926
$\frac{1}{149565}$	0.002282498	0.025901718

Sample Coverage Ratio(SCR): Sample coverage refers to the ratio of our selected neural network training samples to the entire number of samples in the total test area. The first line of the table is an end-to-end test that we conducted under substantial sample coverage. We define it as the reference value 1. Under this benchmark, the second experiment (the second row of the table) expands the range of position and attitude testing, and the total number of samples increases exponentially, but the number of training samples taken remains unchanged. The sample coverage is expressed as 1 in 149,565 of the first experiment. It can be seen that in the case of a reduced proportion of the neural network training set, the model test error has increased. Already stated prompted us to find a practical solution to the visual relationship. On the one hand, the plan should be able to reduce the number of neural network training sets and avoid an exponential increase in the scope of the test. On the other hand, the solution also ensures that the model has a small test error.
Global Mean Square Error(GMSE): On each test model, we randomly select multiple sets of data for performance testing and accuracy analysis. We solve the difference between the actual measured value and the true theoretical value for each data in each information dimension, and square and sum them. The resulting value is divided by 6 to get the mean square error(MSE) value for each data. The MSE values of multiple data are summed and averaged to obtain GMSE. Same as below.
Global Average Error(GAE): We calculate the absolute value of the difference between the actual measured value and the resulting value of each data in each dimension and add them together. On this basis, the results obtained are averaged and finally defined as GAE. Same as below.

TABLE III

Parameters and Test Results Error of Different Networks of Subadded Pictures

Conv Architecture	Num of Parameters	Positive Distance (m)	Lateral Offset (m)	Vertical Offset (m)	Roll Angle (”)	Pitch Angle (”)	Rotation Angle (”)	Global Average Error
SelfConv-5	116.49M	0.003044994	0.003084525	0.012708841	0.012047317	0.000069912	0.000146214	0.005183634
SelfConv-9	45.97M	0.00311657	0.004408261	0.007347916	0.003477437	0.019443595	0.008387517	0.007696883
Resnet-50	50.19M	0.014591278	0.006200716	0.004977294	0.027543208	0.000276501	0.007016631	0.010100938
Resnet-150	84.96M	0.000604705	0.022630479	0.02012717	0.006010174	0.006922322	0.028676191	0.01416184
DenseNet	24.07M	0.007849484	0.029670063	0.013497264	0.022629915	0.006202064	0.040070971	0.019986627

SelfConv-5 is a self-built shallow network architecture with 5 layers of convolutions, and **SelfConv-9** is 9 layers.

On the original basis, the ergodic acquisition of the sample was changed to select a small number of samples within the test range. In the test interval, we randomly picked 50 000 sample images with a different position-and-attitude information for training. These images account for 1/20 000 000 of the total number of theoretical training samples.

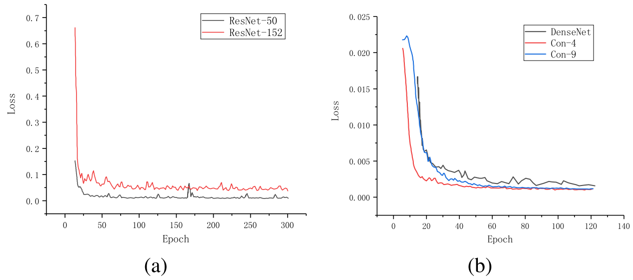
B. Experiments of Subadded Picture

Combined with the theory proposed earlier in this paper, the experiment was tested on subadded picture. In various information dimensions within the test scope, we selected 51 test samples. These test samples equally covered our preferred position and attitude in test ranges. The total number of sample creations in the six information dimensions is 306. Based on the permutation and combination method, we used these images to produce 50 000 subadded pictures. Similarly, the sample coverage is controlled around 1/20 000 000.

We first studied the effect of different convolutional neural network architectures on the recognition accuracy. Five different network architectures have been designed to perform comparative experiments. These neural networks mainly include three categories: self-built shallow convolutional neural networks, ResNet, and DenseNet.

It can be seen that the subadded pictures show a good adaptability to networks of different types and layers. Different depths of ResNet network can achieve final convergence on the verification set, and the convergence results are related to the number of network layers. In the five types of network architectures, we have verified that the loss function of the ResNet–152 network is relatively high. The reason is that too deep convolutional network architecture will extract high level, more abstract posture, and dimensional information. This excessive feature extraction caused a partial loss of accurate information. Therefore, we can infer that a more appropriate network depth will provide good help for feature extraction and information analysis.

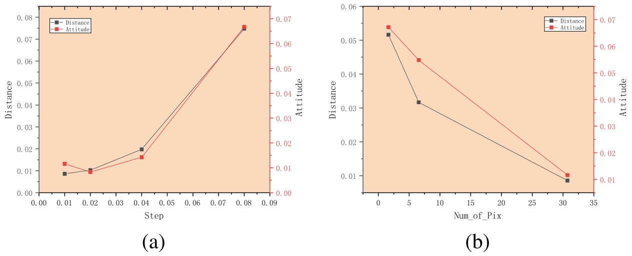
Table III shows the convergence of different networks. It can be seen that the detection of subadded pictures based on convolutional neural networks can obtain useful and fast network convergence results.



(a)

(b)

Fig. 7. Network loss function comparison. Changes in loss with epoch under (a) ResNET-50 and ResNet-152 network architecture and (b) DenseNet, Con-4, and Con9 network architecture.



(a)

(b)

Fig. 8. Effect of image size and stride on accuracy. Visualization of position and attitude measurement errors as a function of (a) step change and (b) Num_of_Pix.

At the same time, we also found that the size of the image input has a significant effect on the detection accuracy. When the size of the input image was 128 × 128, the network training fluctuates, and it was not easy to converge. By adjusting the input image size to 256 × 256, the network achieved fast and effective convergence. Furthermore, we set the picture size to 640 × 480, which reduced the detection error to the minimum of the three. On the other hand, the size of the interval when selecting the samples in each dimension also affected the detection accuracy. Sample preparations with a separation of 0.01 were better than 0.02, 0.04, and 0.08. Figs. 7 and 8 visualize the comparison experiment in the following. We conducted a comparative test on the sample models of four types of unsynchronized amplitudes and three types of network models with different image sizes. It can be seen that as the selection of sample steps continues to increase, both the distance and the angle of the test error have increased. Among them, the model test accuracy of 0.01 and 0.02 is similar. As mentioned earlier, it shows that the measurement model does not distinguish between the features of the two. Therefore, such stride selection has reached the peak of the feature resolution of the image. On the other hand, as the size of the input image continues to increase, the accuracy of its

TABLE IV

IMAGE SIZE IMPACT ON NETWORK TEST RESULTS OF SUBADDED PICTURES

Image Size	Global Mean Square Error	Global Average Error
128*128	0.064712074	0.059377707
256*256	0.010090016	0.04323343
480*640	0.000561968	0.010100938

TABLE V

EFFECT OF STRIDE ON NETWORK TEST RESULTS OF SUBADDED PICTURES

Step Size	Global Mean Square Error	Global Average Error
0.01	0.000561968	0.006060563
0.02	0.000777294	0.00558187
0.04	0.001726988	0.010196057
0.08	0.02024793	0.042469692

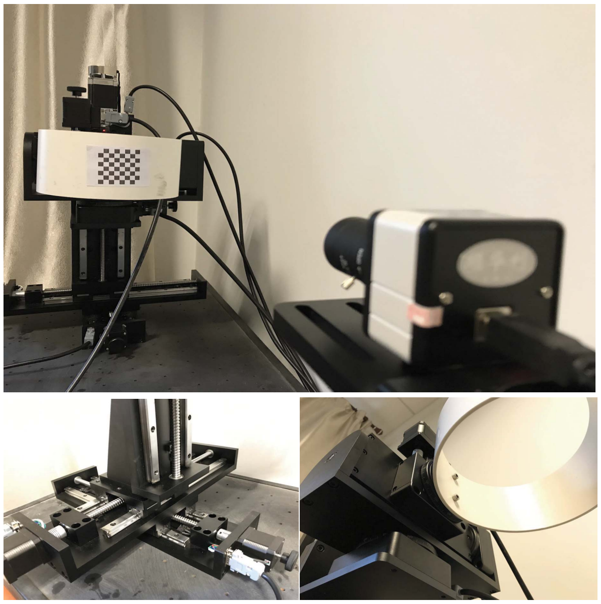


Fig. 9. Physical measurement platform.

model tests continues to rise. On the other hand, as the size of the input image continues to surge, the accuracy of its model tests continues to upturn. We set the acquired image size to 640 × 480 so that we can ensure the test efficiency while guaranteeing the test accuracy. Table IV shows the image size impact on network test results of subadded pictures, and Table V shows the stride effecton network test results of subadded pictures.

C. Experiments for Generating Image Encoder

The subadd theory is the first realization of the visual correspondence relationship in the application of neural networks. This can significantly reduce the amount of training required while ensuring the accuracy of the test. To further realize the effective transformation of the actual test pictures and training pictures, we used the previous theory of generating image encoder to perform the image conversion test.

The picture transformation network is designed as a convolution–deconvolution structure. Based on the previous theory, whether it is the actual shooting picture or subadded picture, all have the same visual correspondence. Therefore, the two can be transformed. We use the real captured image as a network input and the processed image as a network output. The latter corresponds to the type of picture we use when training the network. Based on this, we can achieve the application of the training model in actual testing. In the design process of this network, we found that deep networks like ResNet were more difficult to converge. This problem is related to a large number of parameters to be trained in a deep network. At the same time, after the picture passes through the deep convolutional network, the high-dimensional features are extracted. This high-dimensional feature is more difficult to recover in a deconvolution network. The shallow network can better avoid such problems.

It is worth mentioning that in the specific design process of the convolutional network, we use Leaky rectified linear unit (ReLU) as an activation function. This can effectively avoid the death of neurons caused by large gradients flowing through neurons. At the same time, Batch Normalization (BN) had been introduced into all layers of our network. With the introduction of BN, slight changes in position and attitude information were magnified in the feature map. This dramatically improves the ability of the network to generalize this information.

Based on this, for the first time, we combined the generating image encoder with the theory of visual correspondence and achieved image conversion. At the same time, for the details of network design and parameter control, we also designed several rounds of comparative experiments and recorded relevant data as shown in Table VI.

It can be seen that a flexible shallow network architecture helps to enhance the conversion performance of the image. At the same time, ReLU shows better convergence than tanh. We also note that the number of convolutional cores in each convolutional layer should be set within a reasonable range. This is to ensure that the combination of a fair amount of convolution kernels and shallow convolutional networks has better feature extraction capabilities, and these features can be restored by the deconvolution layer.

Based on the analysis of the network architecture and the details of the parameters, we selected the optimal network for the conversion test. Based on the study of the network architecture and the details of the parameters, we chose the optimal network for the conversion test. The test error of the actual picture is shown in Table VII.

It is worth mentioning that based on the above-mentioned analysis, the network shows good test results, and it can also achieve real-time fast detection. For the equipment used in the experiment, the detection time of our position-and-attitude test model on each picture was approximately 0.075 s.

D. Object Position and Attitude Measurement Experiment in Real Environment

Based on the above-mentioned theoretical analysis and experimental results analysis, object position and attitude

TABLE VI
PARAMETERS AND ARCHITECTURE ADJUSTMENT COMPARATIVE EXPERIMENT OF GENERATING IMAGE ENCODER

No.	Learning Rate	Batch Size	Final Activation	Structure of Conv	Num of Conv	Loss
1	0.0001	16	tanh	7+7	32	0.3651
2	0.003	16	tanh	7+7	32	0.3449
3	0.003	4	tanh	7+7	32	0.1001
4	0.0001	16	tanh	7+7	16	0.3627
5	**0.0001**	**16**	**relu**	**7+7**	**16**	**0.0349**
6	0.0001	16	relu	6+6	16	0.0405

We conducted some comparative experiments on the internal parameters and architecture settings of the Generating Image Encoder network. **Structure of Conv** corresponds to the number of convolutional layers and the number of deconvolution layers in the network. The former number corresponds to the number of convolution layers in the architecture, and the latter number corresponds to the number of deconvolution layers. **Num of Conv** represents the number of convolution kernels owned by each convolution.
The number of convolutional layers is the same to ensure that the characteristics of the network are equivalently transformed. Under the premise that the network can be stabilized, the relu function has better picture conversion performance than the tanh function. Combining with previous theories, the position and attitude features of objects can be identified and transformed by the appropriate number of convolutional neural networks. The number of convolution kernels in each layer should also be set within a reasonable range.

TABLE VII
POSITION AND ATTITUDE TEST ERROR IN GENERATING IMAGE ENCODER

Positive Distance (m)	Lateral Offset (m)	Vertical Offset (m)	Roll Angle (degree)	Pitch Angle (degree)	Rotation Angle (degree)
0.0200	0.0326	0.0075	0.0404	0.0697	0.0336

TABLE VIII
POSITION AND ATTITUDE TEST ERROR IN END-TO-END EXPERIMENTS IN REAL ENVIRONMENTS

Method	Positive Distance (m)	Lateral Offset (m)	Vertical Offset (m)	Roll Angle (”)	Pitch Angle (”)	Rotation Angle (”)
3Dmax Simulation	0.00606	0.00186	0.00149	25.154	19.874	9.758
Actual Situation	**0.00621**	**0.00204**	**0.00179**	**25.365**	**20.265**	**10.005**

measurement based on convolutional neural network proved to have higher accuracy. The relevant conclusions provide adequate support for model training and testing in the actual environment. On this basis, we have built a six-degrees-of-freedom position and attitude measuring device in line with this subject.

The device has a three-degrees-of-freedom position with a minimum movement distance of 0.01 m and a three-degrees-of-freedom attitude with a minimum change angle of 0.01°. The aforementioned provides a guarantee for us to obtain full real training sample acquisition. Furthermore, we placed the actual measurement plane on the measuring device and set a black-and-white checkerboard as a feature marker on the plane. At the same time, we built a visual capture module for capturing the actual image. The physical measurement platform is shown in Fig. 9.

The six-degrees-of-freedom mobile device is controlled by a computer-driven motor. In terms of parameter settings of the device, we control the motor to perform posture rotation and position movement every 0.01 units by running a script.

We use the visual capture module to capture the input image and give the actual position and pose information as a label, making it into a training set. The vision capture module uses a 5–50-mm zoom lens with a field of view of 60°–9.2°. Such lens parameter settings can be effectively matched with the distance between the actual measurement platform and the visual capture device to ensure accurate capture of the feature markers.

Using the images captured by the visual capture device, we trained and tested the neural network. The detection network used is the same as the end-to-end detection network used in Table I, for the comparative analysis between the simulation and the actual measurement. The detection results are shown in Table VIII.

V. CONCLUSION

We proposed an end-to-end position and attitude detection system based on CNN. In the monocular vision system environment, this measurement scheme achieves short-term, high-efficiency, and accurate measurement. Furthermore, based on the visual correspondence theory we have analyzed, we realized the application of subadded pictures in practical engineering situations, which depends on the implementation of generating image encoder. This allows the method to reduce as much unnecessary standard time as possible while maintaining the measurement results. In summary, compared to traditional position and attitude measurement methods, this will effectively promote the further development of intelligent precision measurement in theory and engineering. We will explore the influence of the proportion of information in each dimension on subadded pictures and try to incorporate sufficient relevant information besides position and attitude. At the same time, more effective signature extraction and pretreatment programs will be investigated. Earlier stated will improve the system's accuracy and universality in engineering applications.

References

[1] B. Jiang, J. Yang, Z. Lv, and H. Song, "Wearable vision assistance system based on binocular sensors for visually impaired users," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1375–1383, Apr. 2019.

[2] P. Luo, L. Lin, and X. Liu, "Learning compositional shape models of multiple distance metrics by information projection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 7, pp. 1417–1428, Jul. 2016.

[3] A.-M. Zou and K. D. Kumar, "Neural network-based distributed attitude coordination control for spacecraft formation flying with input saturation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1155–1162, Jul. 2012.

[4] P. Jasiobedzki, S. Se, T. Pan, M. Umasuthan, and M. Greenspan, "Autonomous satellite rendezvous and docking using LIDAR and model based vision," *Proc. SPIE*, vol. 5798, May 2005, pp. 54–66.

[5] J. Galante *et al.*, "Pose measurement performance of the argon relative navigation sensor suite in simulated-flight conditions," in *Proc. AIAA Guid., Navigat. Control Conf.*, 2012, p. 4927.

[6] J. M. Kelsey, J. Byrne, M. Cosgrove, S. Seereeram, and R. K. Mehra, "Vision-based relative pose estimation for autonomous rendezvous and docking," in *Proc. IEEE Aerosp. Conf.*, Mar. 2006, p. 20.

[7] M. Balch and D. Tandy, "A pose and position measurement system for the hubble space telescope servicing mission," *Proc. SPIE*, vol. 6555, May 2007, Art. no. 65550F.

[8] S. Augenstein, *Monocular Pose Shape Estimation Moving Targets, For Auto. Rendezvous Docking*. Stanford, CA, USA: Stanford Univ., 2011.

[9] T. Boge, H. Benninghoff, M. Zebenay, and F. Rems, "Using robots for advanced rendezvous and docking simulation," in *Proc. Workshop Simulation EGSE Space Programmes (SESP)*, 2012.

[10] R. T. Howard, T. C. Bryan, M. L. Book, and J. L. Jackson, "Active sensor system for automatic rendezvous and docking," *Proc. SPIE*, vol. 3065, Aug. 1997, pp. 106–116.

[11] R. T. Howard, H. J. Cole, J. L. Jackson, G. W. Kamerman, and D. K. Fronek, "Automatic rendezvous and docking system test and evaluation," *Proc. SPIE*, vol. 3065, pp. 131–140, Aug. 1997.

[12] R. T. Howard, A. S. Johnston, T. C. Bryan, and M. L. Book, "Simulation and ground testing with the AVGS," *Proc. SPIE*, vol. 5799, May 2005, pp. 56–66.

[13] N. A. S. Johnston, R. T. Howard, and D. W. Watson, "X-Ray calibration facility/advanced video guidance sensor test," NTRS, Huntsville, AL, USA, Tech. Rep. NASA/TM-2004-213393, M-1122, 2004.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[15] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[16] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[17] A. Heydari and S. N. Balakrishnan, "Finite-horizon control-constrained nonlinear optimal control using single network adaptive critics," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 1, pp. 145–157, Jan. 2013.

[18] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*. Natick, MA, USA: Gatesmark, 2009.

[19] R. Szeliski, *Computer Vision: Algorithms and Applications*. Berlin, Germany: Springer, 2010.

[20] S. Gold, C.-P. Lu, A. Rangarajan, S. Pappu, and E. Mjolsness, "New algorithms for 2D and 3D point matching: Pose estimation and correspondence," in *Proc. Adv. Neural Inf. Process. Syst.*, 1995, pp. 957–964.

[21] G.-Q. Wei and S. D. Ma, "Implicit and explicit camera calibration: Theory and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 5, pp. 469–480, May 1994.

[22] Z. Zhang, "Camera calibration with one-dimensional objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 7, pp. 892–899, Jul. 2004.

[23] E. Hecht, *Optics*. London, U.K.: Pearson Education, 2016.

[24] A. F. Heaton and R. T. Howard, "POSE algorithms for automated docking," *Proc. SPIE*, vol. 8044, May 2011, Art. no. 80440T.

[25] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2147–2156.

[26] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," 2017, *arXiv:1711.00199*. [Online]. Available: https://arxiv.org/abs/1711.00199

[27] Z. Li, W. Yuan, Y. Chen, F. Ke, X. Chu, and C. L. P. Chen, "Neural-dynamic optimization-based model predictive control for tracking and formation of nonholonomic multirobot systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 12, pp. 6113–6122, Dec. 2018.

[28] G. Cheng, P. Zhou, and J. Han, "Duplex metric learning for image set classification," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 281–292, Jan. 2018.

[29] J. Han, G. Cheng, Z. Li, and D. Zhang, "A unified metric learning-based framework for co-saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2473–2483, Oct. 2017.

[30] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[31] J. Han, R. Quan, D. Zhang, and F. Nie, "Robust object co-segmentation using background prior," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1639–1651, Apr. 2018.

[32] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[33] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2014, pp. 818–833.

[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[35] C. Szegedy *et al.*, "Going deeper with convolutions" in *Proc. CVPR*, Jun. 2015, pp. 1–9.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[37] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. IEEE CVPR*, vol. 1, no. 2, Jun. 2017, p. 3.

[38] Y. Hao, F. Zhu, and J. Ou, "Three-dimensional visual methods for object pose measurement," *Proc. SPIE*, vol. 4553, pp. 78–83, Sep. 2001.

[39] J. J. Craig, "Introduction to robotics: Mechanics and control," Dept. Elect. Comput. Eng., Univ. California at Santa Barbara, Santa Barbara, CA, USA, 1955.

[40] T. Imaida, Y. Yokokohji, T. Doi, M. Oda, and T. Yoshikawa, "Ground-space bilateral teleoperation of ETS-VII robot arm by direct bilateral coupling under 7-s time delay condition," *IEEE Trans. Robot. Autom.*, vol. 20, no. 3, pp. 499–511, Jun. 2004.

[41] R. T. Howard *et al.*, "The advanced video guidance sensor: Orbital express and the next generation," in *Proc. AIP Conf.*, 2008, p. 717.

[42] R. T. Howard and T. C. Bryan, "DART AVGS flight results," *Proc SPIE*, vol. 6555, May 2007, Art. no. 65550L.

[43] J. Lee, C. Carrington, S. Spencer, T. Bryan, R. Howard, and J. Johnson, "Next generation advanced video guidance sensor: Low risk rendezvous and docking sensor," in *Proc. AIAA SPACE Conf. Expo.*, 2008, p. 7838.

[44] S. Shi, L. Yang, J. Lin, Y. Ren, S. Guo, and J. Zhu, "Omnidirectional angle constraint based dynamic six-degree-of-freedom measurement for spacecraft rendezvous and docking simulation," *Meas. Sci. Technol.*, vol. 29, no. 4, 2018, Art. no. 045005.

[45] F. Chollet *et al.*, (2015). *Keras*. [Online]. Available: https://github.com/fchollet/keras

[46] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.